1К

УДК 631.3.07

Олег Клічук
Чернівецький національний університет імені Юрія Федьковича, Україна

Особливості інтелектуальних методів кластеризації у реляційних базах даних

У статті досліджується методика проведення кластерного аналізу у реляційних базах даних на основі функціонального програмування. Описано функції на основі рекурентних нейронних мереж та самоорганізованих карт Коханена для проведення кластерного аналізу.

Сучасні досягнення в інформаційних технологіях конкретизуються у практичному їх втіленні, що підтверджує прикладне значення і застосування штучного інтелекту. Прогрес у технологічних засобах очевидний. Проте є ряд проблем, які потребують детальнішого аналізу та уточнення, що обумовлено їх практичним використанням. До сфери таких проблем можна віднести використання інтелектуальних методів кластеризації даних у інформаційних системах.

Проблеми кластерного аналізу розглянуто у Манделя [1], [2]. Ці дослідження переважно теоретичного характеру, а саме: містять прикладне значення кластеризації даних, основні алгоритми проведення кластеризації.

Інтелектуальні методи обчислень набули розвитку в сучасній науковій думці, зокрема у працях [3], [4]. Основою таких обчислень є нейронні мережі, генетичні алгоритми та інші методи.

Окремі дослідження у методах обробки реляційних баз даних є, але їх практичне застосування не має системного характеру. У розвитку сучасних інформаційних систем спостерігається активний пошук нових методів обробки баз даних, тому що в основі кожної інформаційної системи присутня реляційна база даних.

Класифікація об’єктів по осмислених групах – кластеризація – є важливою процедурою у сфері економічних, соціологічних, психологічних досліджень і фундаментальним процесом наукової практики, тому що системи класифікацій містять поняття, необхідні для розробки теорій у науці. Отже, **метою даного дослідження** є пошук та аналіз методів обробки даних у реляційних базах даних. Це полегшить звичайним користувачам використання потужних інтелектуальних методів обробки даних.

Кластерний аналіз – загальна назва множини обчислювальних процедур, які використовуються при створенні класифікації. У результаті роботи з процедурами утворюються «кластери» або групи дуже подібних об’єктів. Більш точно, кластерний метод – багатомірна статистична процедура, яка виконує збір даних, які містять інформацію про вибірку об’єктів, а після – упорядковує об’єкти у порівняно однорідні групи.

Кількісна оцінка подібності побудована на понятті метрики. При цьому підході кожен об’єкт множини позначається точками координатного простору, при цьому помічені подібності і відмінності між точками знаходяться у відповідності з метричними відстанями між ними. Розмірність простору визначається кількістю змінних, які використовуються для опису подій.

Використовують такі стандартні властивості критеріїв, яким повинні відповідати міри подібності, щоб бути метрикою:
– симетрія;

– нерівність трикутника;
– відмінність неоднакових об’єктів;
– відсутність відмінностей ідентичних об’єктів.

Найбільш широковживаними функціями відстаней між об’єктами є:
– евклідова відстань:

$$d_2\,(X_i\,,X_j\,) = [\ \sum_{k=1}^{p}\ (x_{ki} - x_{kj}\,)^2\ ]^{1/2}; \tag{1}$$

– l_1 – норма:

$$d_1\,(X_i\,,X_j\,) = [\ \sum_{k=1}^{p}\ |\,x_{ki} - x_{kj}\,|\ ]; \tag{2}$$

– супремум-норма:

$$d_\infty\,(X_i\,,X_j\,) = \sup\ \{\ |\,x_{ki} - x_{kj}\,|\ \},\ k = 1, 2, 3, \ldots p; \tag{3}$$

– l_p – норма:

$$d_p\,(X_i\,,X_j\,) = [\ \sum_{k=1}^{p}\ |\,x_{ki} - x_{kj}\,|^p\ ]^{1/p}; \tag{4}$$

– віддаль Махаланобіса:

$$D^2\,(X_i\,,X_j\,) = (X_i - X_j\,)^T\,W^{-1}\,(X_i - X_j\,). \tag{5}$$

Проблема вимірювання близькості об’єктів постійно виникає при будь-яких тлумаченнях кластерів та різних методах класифікації. Основними труднощами при цьому є неоднозначність вибору способу нормування і знаходження віддалі між об’єктами.

Незважаючи на важливість евклідової та інших метрик, вони мають значні недоліки, з яких найбільш суттєвий полягає у тому, що оцінка подібності дуже залежить від відмінностей у зсувах даних. Змінні, у яких наявні одночасно великі абсолютні значення і стандартні відхилення, можуть подавити вплив змінних з меншими абсолютними значеннями та стандартними відхиленнями. Більш того, метричні відстані змінюються під дією перетворення шкали вимірювання змінних, при яких не зберігається ранжування за евклідовою віддаллю.

Вибір змінних у кластерному аналізі є одним з найбільш важливих кроків у процесі дослідження, але і одним з найменш розроблених. Основна проблема полягає у тому, щоб знайти ту сукупність змінних, яка оптимальним чином відображає поняття подібності та описує об’єкти. В ідеалі змінні повинні вибиратися у відповідності з чітко сформульованою теорією, яка лежить в основі класифікації. В інформаційних системах такими змінними є властивості об’єктів. Вони можуть мати як кількісне, так і якісне значення. Тому при проведенні обчислень виникає необхідність перетворення якісних даних у числові.

Також у більшості видів аналізу дані, звичайно, підлягають нормуванню певним способом. У тому випадку якщо дані виміряні у різних масштабах, нормування, звичайно, проводиться таким чином, щоб середнє арифметичне дорівнювало нулю, а дисперсія – одиниці.

Нормування являє собою перехід до певного однозначного опису для всіх ознак, до введення нової умовної одиниці вимірювання, яка допускає формальне співставлення об’єктів. Такі процедури у реляційних базах даних можна проводити за допомогою підсумкових операцій, декартового добутку та обчислень на основі функціонального програмування. Найбільш поширеним способом нормування властивостей є:
– обчислення підсумковою функцією середнього арифметичного $\vec{x}$;
– обчислення підсумковою функцією середньоквадратичного відхилення □;

– обчислення підсумковою функцією максимального значення x_{max};
– обчислення підсумковою функцією мінімального значення x_{min};
– проведення операції декартового множення з відношеннями даних, максимального, мінімального, середнього арифметичного значення і середньоквадратичного відхилення;
– знаходження нормованого значення за допомогою функції нормування: $x' =(x - \vec{x}\,) /$ □ $/ (x_{max} - x_{min}\,)$

Такий підхід забезпечує просту та зручну організацію додаткових засобів обробки інформації у готових базах даних.

Незважаючи на відсутність чіткого означення, кластери володіють деякими властивостями, найважливішими з яких є густина, дисперсія, розміри, форма і відокремленість.

Основні кластерні методи можна поділити на такі групи:
– ієрархічні агломеративні методи;
– ієрархічні дивизімні методи;
– ітеративні методи групування;
– методи пошуку модальних значень густини;
– факторні методи;
– методи згущень;
– методи, які використовують теорію графів.

Ці групи методів відповідають різним підходам до створення кластерів, і використання різних методів до одних і тих же даних може привести до суттєво відмінних результатів. У конкретних галузях науки найчастіше застосовують характерні групи методов кластеризації. Так, ієрархічні агломеративні методи частіше за все використовуються у біології, тоді як факторні аналітичні методи з великим успіхом використовуються у психології. При виборі методу кластеризації необхідно враховувати відповідність цього методу до очікуваного характеру класифікації, використаних ознак і міри подібності. Найбільш відомими групами кластерних методів, які використовуються у соціальних науках, є ієрархічні агломеративні, ієрархічні дивизімні і факторні.

Основними причинами розробки та використання спеціальних методів статистичного аналізу багатомірних даних є необхідність розуміння закономірностей функціонування недостатньо вивчених складних соціально-економічних процесів і явищ, а також використаннях цих методів як інструменту управління, який призначений для аналізу багатомірних реальних, швидкозмінних ситуацій. Основою сучасних інформаційних систем управління є реляційні бази даних.

При обробці реляційних баз даних необхідно враховувати методику зберігання та доступу до інформації у них. Вона полягає у відокремленості записів у відношенні і проведенні обчислень поступово за кожним записом. Такі обмеження вимагають проведення пошуку специфічних методів обробки. Тобто з одного боку – простий метод доступу до даних у відношеннях, з іншого – обмеженість у використанні даних з різних записів відношення. Перераховані вище методи кластеризації вимагають одночасного порівняння кількох об’єктів.

У звязку з відсутністю залежностей між записами реляційних баз даних та обмеженістю операцій реляційної алгебри сукупний аналіз інформації повинен бути забезпечений функцією проведення кластеризації.

Одним із перспективних методів є використання функціонального програмування. Суть його полягає у проведенні обчислень на основі тільки функцій [5]. У даному випадку аргументами функції є значення полів.

Найпростішими функціями пошуку кластера можуть бути функції, які містять дані про центр розміщення кластера та його форму. Послідовний підхід у побудові алгоритму кластеризації полягає в явному формулюванні певного цільового функціонала якості з наступною його мінімізацією. Найпростішим функціоналом якості є сумарна віддаль по всіх зразках до кожного об'єкта до центра найближчого від нього кластера у вибраній метриці. Цей функціонал можна визначити для умови, коли число кластерів відомо наперед. Змінними пошуку є координати центрів кластерів. Мінімізація функціонала якості повинна проводитися за всіма можливими перестановками об'єктів по кластерах. Це і визначає фундаментальну складність задачі кластеризації.

Більш потужні функції кластеризації можна побудувати на основі нейронних мереж. Для розв'язування будь-якої задачі з використанням штучних нейронних мереж необхідно спочатку зпроектувати структуру мережі, адекватну поставленій задачі. Це передбачає вибір кількості шарів мережі і нейронів у кожному шарі, а також визначення необхідних зв'язків між шарами.

Підбір кількості нейронів у вхідному шарі обумовлений розмірністю вхідного вектора, у даному випадку – кількістю полів відношення. Подібна ситуація із вихідним шаром: у випадку з кластеризацією це, як правило, одне значення. Складним питанням залишається підбір кількості прихованих шарів і кількості нейронів у кожному з них. Теоретичний розв'язок цієї задачі у сенсі умови достатності був запропонований математиками, які вивчають апроксимацію функцій декількох змінних.

У досліджуваному напрямку варто звернути увагу на окрему групу нейронних мереж зі зворотнім зв'язком між різними шарами нейронів. Це – так звані рекурентні мережі. Їх загальна ознака полягає у передачі сигналів із вихідного, або схованого, шару у вхідний шар.

Основна особливість, яка виділяє ці мережі серед інших нейронних мереж, – динамічна залежність на кожному етапі функціонування. Зміна стану одного нейрона відображається на всій мережі внаслідок зворотнього зв'язку типу «один до багатьох». У мережі виникає певний перехідний процес, який завершується формуванням стійкого стану, який відрізняється у загальному випадку від попереднього. Ці стани можна позначати кластерами, які відповідають таким множинам об'єктів.

Якщо функцію активації нейрона позначити $f(u)$, де u – це зважена сума його збудження, то стан нейрона можна визначити вихідним сигналом $y_i = f(u_i) = f\left(\sum_{j=1}^{N} w_{ij} x_j\right)$.

Беручи до уваги те, що при зворотньому зв'язку типу «один до багатьох» роль збуджених імпульсів для нейрона відіграють вихідні сигнали інших нейронів, зміну його стану можна описати системою диференціальних нелінійних рівнянь:

$$\tau_i \frac{du_i}{dt} = \sum_{j=1, j\neq i}^{N} w_{ij} f(u_j) - u_i - b_i \tag{6}$$

для $i = 1, 2, \ldots, N$, де b_i є пороговим значенням, заданим зовнішнім джерелом. Коефіцієнт τ_i – числова константа, яка описує динамічний стан. Стан нейрона розраховується розв'язком такого рівняння, як $y_i = f(u)$. При певному рівні збудження нейронів, який описується значеннями їх вихідних сигналів y_i, з рекурентною мережею можна співставити енергетичну функцію Ляпунова:

$$E = -\frac{1}{2}\sum_{j}\sum_{i, i\neq j} w_{ij} y_i y_j + \sum_{i=1}^{N} \frac{1}{R_i}\int_0^{x_i} f_i^{-1}(y_i)\,dy_i + \sum_{i=1}^{N} b_i y_i . \tag{7}$$

Вона пов'язана з кожним збудженим станом мережі і має тенденцію зменшуватися з часом. Зміна стану кожного нейрона ініціюється зміною енергетичного стану всієї мережі у напрямі мінімуму її енергії, аж до його досягнення. Звичайно, існує багато локальних мінімумів, кожен з яких являє собою один зі станів системи, який визначається структурою мережі. У просторі станів локальні енергетичні мінімуми енергії подані точками стабільності, які називаються атракторами через тяжіння до них найближчого оточення.

Однією з найбільш досліджених рекурентних мереж є мережа Хопфілда. Узагальнена структура цієї мережі являє собою систему з безпосереднім зворотнім зв'язком виходу з входом. Характерною особливістю такої системи є те, що вихідні сигнали нейронів є одночасно вхідними сигналами мережі. У класичній системі Хопфілда відсутні зв'язки нейрона з власним виходом, це полегшує процес її налаштування.

Процес навчання мережі формує зони притягання (кластери) точок рівноваги. Найчастіше нейрони мережі Хопфілда мають функцію активації типу ***signum*** зі значеннями ±1. Це означає, що вихідний сигнал ***i***-го нейрона визначається функцією

$$y_i = sgn(\sum_{j=0}^{N} w_{ij}x_j + b_i), \tag{8}$$

де N – кількість нейронів; w_{ij} – матриця синоптичних зв'язків, b_i – коефіцієнти вектора зсуву.

Механізм модифікації синаптичних зв'язків запропонований математичною моделлю Хебба. Для опису правила Хебба у математичних термінах розглянемо синаптичний зв'язок нейрона k з передсинаптичним та післясинаптичним сигналами x_j та y_k. Зміну величини синаптичного зв'язку у момент часу n можна записати у вигляді:

$$\Delta w_{kj}(n) = F(y_k(n), x_j(n)), \tag{9}$$

де $F(\,)$ – певна функція, яка залежить від передсинаптичних та післясинаптичних сигналів.

Ця формула може бути записана у наступній формі:

$$\Delta w_{kj}(n) = \eta\, y_k(n)\, x_j(n), \tag{10}$$

де η – додатня константа, яка визначає швидкість навчання.

Для навчання без учителя можна використати правило конкурентного навчання. Наприклад, можна використати нейронну мережу, яка складається з двох шарів – вхідного та вихідного. Вхідний шар отримує доступні дані. Вихідний складається з нейронів, які конкурують між собою за право відклику на ознаки, які містяться у вхідних даних. У найпростішому випадку нейронна мережа працює за принципом «переможець отримує все». При такій стратегії нейрон з найбільшим сумарним вхідним сигналом «перемагає» у змаганні і переходить в активний стан, при цьому всі інші нейрони відключаються.

Для проведення кластеризації можна використовувати функції, які побудовані на основі алгоритмів самоорганізованого навчання. Метою цих алгоритмів є виявлення у множині вхідних даних суттєвих ознак об'єктів. Для цього алгоритм реалізує правила локальної природи, що дозволяє проводити навчання обчислення відображеного вхідного сигналу на вихідний з потрібними властивостями. Під терміном «локальний» розуміють заміну синаптичних ваг тільки безпосередніми сусідами цього нейрона. Моделі мереж, які навчаються на основі принципів самоорганізації, відображають властивості нейробіологічних структур. Архітектура самоорганізованих систем може

приймати багато різних форм. Процес навчання полягає у періодичній зміні синаптичних ваг всіх зв’язків у системі у відповідь на подачу вхідних зразків у відповідності з призначеними правилами для отримання відповідної конфігурації системи.

Алгоритм, відповідальний за формування самоорганізуючих карт, починається з ініціалізації синаптичних ваг мережі. Звичайно це відбувається із призначенням синаптичним вагам малих значень, які сформовані генератором випадкових чисел. При такому формуванні карта ознак початково не має якого-небудь порядку ознак. Після коректної ініціалізації мережі для формування карти самоорганізації запускаються три наступні основні процеси:

– конкуренція (competition) – для кожного вхідного зразка нейрони мережі обчислюють відносні значення дискримінантної функції, ця функція є основою конкуренції серед нейронів;
– кооперація (cooperation) – нейрон, який переміг, визначає простір положення топологічного околу нейронів, який забезпечує базис для кооперації між цими нейронами;
– синаптична адаптація (synaptic adaptation) – цей механізм дозволяє збудженим нейронам збільшувати власні значення дискримінантних функцій по відношенню до вхідних образів за допомогою відповідних коректувань синаптичних ваг, зміна проводиться таким чином, щоб відклик нейрона-переможця на наступні аналогічніні приклади посилювався.

Математична модель процесу конкуренції наступна. Нехай m – розмірність вхідного простору, а вхідний вектор вибирається з цього вхідного простору випадково і позначається як:

$$\boldsymbol{X} = [\, x_1, x_2, \dots, x_m \,]^T.$$

Вектор синаптичних ваг кожного з нейронів мережі має таку саму розмірність, що і вхідний простір. Позначимо синаптичну вагу нейрона j:

$$\boldsymbol{W} = [\, w_{j1}, w_{j2}, \dots, w_{jm} \,]^T, j = 1, 2, \dots, l,$$

де l – загальна кількість нейронів мережі. Для того щоб підібрати найкращий вектор W_j, який відповідає вхідному вектору $\boldsymbol{X}$, необхідно порівняти скалярні добутки $\boldsymbol{W_j}^T \times \boldsymbol{X}$ для $j = 1, 2, \dots, l$ і вибрати найбільше значення. Таким чином, вибравши нейрон з найбільшим скалярним добутком, ми в результаті визначаємо місцезнаходження, яке повинне стати центром топологічного околу збудженого нейрона.

Для проведення операції кооперації необхідно визначити окіл збудженого нейрона. Типовим прикладом обчислення цієї відстані є функція Гаусса:

$$h_{j,i(x)} = exp(-d^2_{j,i} / (2\sigma^2)), \tag{11}$$

де $d_{j,i}$ – латеральна віддаль (lateral distance) між нейроном-переможцем (i) та вторинно збудженим нейроном (j); σ – параметр, який визначає рівень, до якого нейрони з топологічного околу нейрона-переможця приймають участь у процесі навчання.

Для того щоб мережа могла самоорганізовуватися, вектор синаптичних ваг нейрона повинен змінюватися у відповідності до вхідного вектора. Основна проблема полягає у тому, як зміна повинна проходити. Враховуючи правило Хебба про те, що синаптична вага повинна підсилюватися при одночасному виникненні предсинаптичної і постсинаптичної активності та складової забування (forgetting term), зміна синаптичних ваг має вигляд:

$$\Delta \boldsymbol{w}_j(n) = \eta(n)\, h_{j,i(x)}(n)\, (\boldsymbol{x} - \boldsymbol{w}_j(n)), \tag{12}$$

де η – параметр швидкості навчання (1earning-rate parameter) алгоритму. Цей вираз викристовується для всіх нейронів решітки, які лежать у топологичному околі нейрона-переможця. Він має ефект переміщення вектора синаптичних ваг нейрона-переможця у бік вхідного вектора.

Висновки

У роботі досліджено методику кластеризації реляційних баз даних на основі функціонального програмування.

У ролі засобів функціонального програмування запропоновано функції на основі рекурентних нейронних мереж та самоорганізаційних карт Коханена.

Література

1. Мандель И.Д. Кластерный анализ / Мандель И.Д. – М. : Финансы и статистика, 1988. – 176 с.
2. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл ; [пер. с англ. Е.З. Демиденко ; под ред. А.Я. Боярского]. – М. : Статистика, 1977. – 128 с.
3. Осовский С. Нейронные сети для обработки информации / С. Осовский ; [пер. с польского И.Д. Рудинского]. – М. : Финансы и статистика, 2002. – 344 с.
4. Бэстенс Д.-Э. Нейронные сети и финансовые рынки: принятие решений в торговых операциях / Бэстенс Д.-Э., ван ден Берг В.-М., Вуд Д. – М. : ТВП, 1997. – 236 с.
5. Клічук О. Обробка реляційних баз даних засобами функціонального програмування / О. Клічук // Искусственный интеллект. – 2009. – № 2. – С. 57-63.

Олег Кличук

Особенности интеллектуальных методов кластеризации в релятивных базах данных
В статье исследована методика проведения кластерного анализа в реляционных базах данных с использованием функционального программирования. Описаны функции на основе реккурентных нейронных сетей и самоорганизующихся карт Коханена для проведения кластерного анализа.

Стаття надійшла до редакції 19.01.2010.